Apr 9th, 12:00 AM

Preserving Scientific Research Data at Middlebury College

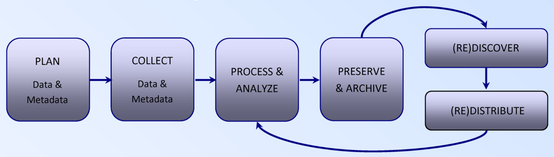
Wendy Shook
Middlebury College

Preserving Scientific Research Data at Middlebury College

Wendy Shook
Middlebury College

Problem

Middlebury College is a small liberal arts college with an active and growing scientific research community. Recognizing that data are a valuable resource, research funding agencies require preservation and accessibility of data and its metadata. Open data is also becoming a strong motivation for preservation and access. For these to happen, data should be curated through the whole data lifecycle.



There are large discipline specific repositories, but not all data, including a percentage of that created by Middlebury researchers, fit the ingestion criteria. For these data sets, however large or small, Middlebury College Library is working towards implementation of a local science data repository to preserve and make accessible research products for the long term.

Planning

Planning for such a repository was divided into five main stages.

- Survey and assess faculty needs for on-campus data storage.
- Map the elements of a sustainable data repository.
- Research potential platforms.
- Draft policies and procedures, adopting standards and best practices when possible.
- Test and assess prototype(s) of most suitable platform(s).

	CONTENTdm	DataFlow	Dataverse	DSpace	Fedora Commons	Islandora	Omeka
initial purchase	fee	$0	$0	$0	$0	$0	$0
annual maintenance	fee	$0	$0	$0	$0	$0	$0
source code	closed	open	open	open	open	open	open
license	proprietary	MIT	Apache	BSD	Apache	GPL	GPL
initial release	2001	pending	2006	2002	2003	2009	2008
developer	Washington	Oxford	Harvard	MIT/HP	Cornell/Virginia	Prince Edward Island	George Mason
maintainer	OCLC	Oxford	Harvard	DuraSpace	DuraSpace	Prince Edward Island	George Mason
status	active	active	active	active	active	active	active
age	mature	young	intermediate	mature	mature	intermediate	intermediate
version	6	0.3b	3.2	1.0	3.6.2	12.2.0	1.53
operating system	Windows/Linux	Linux	Linux/Mac	Linux/Mac/Windows	Linux/Mac/Windows	Linux/Mac/Windows	Linux
application server	Apache/IIS		Glassfish	Tomcat	Tomcat	Tomcat	Apache
database	proprietary		PostgreSQL	PostgreSQL/Oracle	various	various	MySQL
code	PHP	Python	Java	Java	Java	Java	PHP
other						Fedora Commons/Drupal	

Pilots

First Pilot: CONTENTdm

CONTENTdm is a proprietary institutional repository platform that is optimized for library collections, developed at the University of Washington then acquired and maintained by OCLC. It was chosen as the platform for our first pilot because it was available and well supported at Middlebury. While well suited to library needs and offering vendor support, it lacks flexibility needed to support a robust data repository.

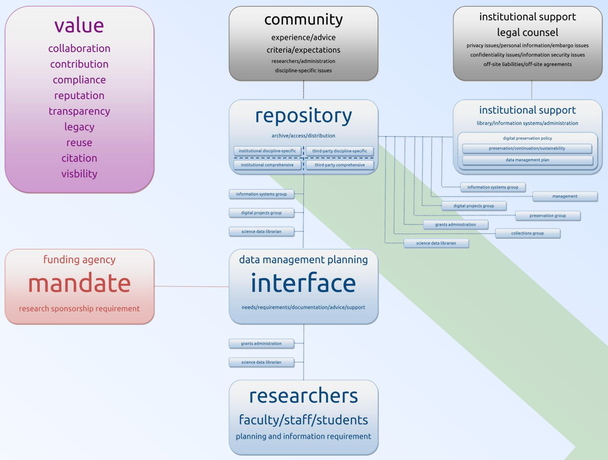
Second Pilot: Islandora

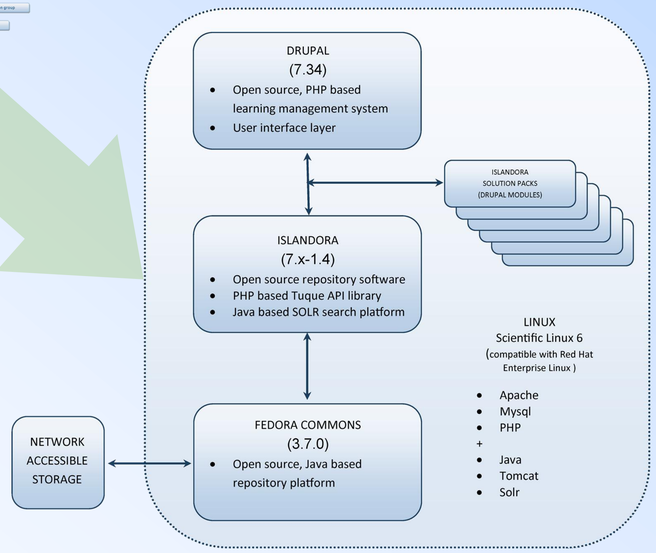
Islandora is an open source digital asset management system developed by the University of Prince Edward Island that uses Fedora Commons and Drupal. Islandora itself is a Drupal module that provides a link between the Drupal CMS user interface and the Fedora repository structure. Additionally, Islandora optionally uses the Java-based Solr search and discovery platform.

Islandora currently enjoys several advantages over CONTENTdm. It allows nested collections and also has the ability to set independent metadata and security requirements for individual collections and sub-collections, including access and embargo controls. Islandora is extensible, flexible, and platform-independent. Lastly, Islandora has a large and active development and support community.

Flexibility comes with a cost. Implementation of Islandora may require in-house skills for installation and maintenance, which can overwhelm a small support staff (although support can be outsourced to third parties).

With greater flexibility also comes the burden of creating sound, sustainable policies for discipline specific metadata, naming conventions, intellectual property concerns, access, and security. Islandora has tools to simplify these processes, but policy is developed at the institutional level.





Next Steps

Data@Middlebury, or D@M, has been commissioned locally as a functional proof-of-concept prototype, and we are assessing and optimizing functionality and integration.

The current instance was developed as a standalone demonstration piece. The next step is to recreate it on an institutional level, incorporating it into the institutional Drupal instance, with integrated networked storage. There is also interest from the Digital Arts Initiative to integrate Omeka with the Fedora Commons instance – a relatively simple customization with an Omeka module. Policies are being evaluated, with input from faculty. This will necessarily lead to iteration of the policies and procedures developed for the first pilot.

A concern that needs to be addressed is the risk of becoming an isolated data island; researchers know, or can readily find, data at large, discipline specific repositories, but discovery at a smaller institution is less likely. DOIs may not be enough, and exciting research could be done into cross-listing in meta-repositories.

Resources

CONTENTdm: http://www.contentdm.org/
Drupal: https://www.drupal.org/
Fedora Commons: http://fedorarepository.org/
Islandora: http://islandora.ca/
Omeka: http://omeka.org/